A. van Strien, J. Pannekoek, W. Hagemeijer & T. Verstrael, 2004 - A loglinear Poisson regression method to analyse bird monitoring data. In: Anselin, A. (ed.) Bird Numbers 1995, Proceedings of the International Conference and 13th Meeting of the European Bird Census Council, Pärnu, Estonia. *Bird Census News* 13 (2000):33-39

A LOGLINEAR POISSON REGRESSION METHOD TO ANALYSE BIRD MONITORING DATA

A. van Strien[1], J. Pannekoek[1], W. Hagemeijer[2]& T. Verstrael[1]

ABSTRACT. A loglinear Poisson regression method has been developed to analyse time series of count data. The method produces yearly indices and trend estimates. It is also capable of testing the effects of covariants on the changes so that the impact of human activities on changes can be investigated. The method can also deal with several difficulties inherent to monitoring data, especially missing values, over- and undersampling of particular strata, serial correlation and deviations from Poisson distribution. A user-friendly computer program, called TRIM, performs application of the method.

[1] Statistics Netherlands, P.O. Box 4000,
2270 JM Voorburg, The Netherlands
[2] SOVON Dutch Centre for Field Ornithology, Rijksstraatweg 178,
6573 DG Beek-Ubbergen, The Netherlands

INTRODUCTION

The increasing time span of bird monitoring programs yields a large amount of valuable count data. Usually these data concern yearly counts of individuals, breeding pairs or territories, which are being transformed into yearly indices by applying a statistical method. But more attention is needed for the statistical tools applied. One reason is that a number of currently used methods is not well suited to deal with the problems inherent in the time series (Ter Braak et al., 1994). A second reason is that the information need is growing, especially the need to test for trends. Here, we discuss these reasons in more detail and then report on a statistical method that can cope with the problems in the data and with the increased information need.

PROBLEMS IN THE DATA

Total counts across all sample sites from one year to another produce the time course of the numbers of breeding pairs. These counts are usually represented as indices, using the first year as a base year. But for all sorts of reasons missing values in the time series may easily enter the data set. In the Dutch Common Breeding Bird Census, in which Statistics Netherlands co-operates with SOVON, the Dutch Centre for Field Ornithology, up to 60 % of the counts can be missing. The occurrence of missing values hampers the production of index numbers because simple comparisons of total numbers between years give misleading results, since such comparisons reflect not only changes in the number of individuals, but changes in the pattern of missing values as well. In case this problem is not solved properly, it can even lead to artificial trends, a phenomenon known as 'random walk' (Crawford, 1991).

A second problem in monitoring programs is the oversampling of particular areas and the undersampling of others. This happens when volunteers prefer to count sites in more attractive areas, in the Netherlands for example dunes, heathland and marshes, instead of other areas like conifer plantations and urban areas. This also happens when particular areas are oversampled to enable the assessment of the effects of human activities, such as management. As a result, no proper national indices can be produced straightforwardly, because the changes are often not similar in all area types.

Furthermore, count data are not normally distributed. Transformation of counts is often unsatisfactory, especially in case of many zero counts. Transformation of counts can be circumvented by the use of loglinear regression (a form of GLM, Generalized Linear Models, see McCullagh & Nelder (1989) and Dobson (1991)). Loglinear regression is based on the assumption of independent Poisson distribution for the counts. A complication is that such an assumption is likely to be violated for counts of birds because the variance is often larger than expected for a Poisson distribution, especially when they occur in groups. This phenomenon is called overdispersion.

Finally, counts in a particular year will also depend on the counts in the year before, a phenomenon called serial correlation. This should be taken into account while estimating standard errors and testing trends.

INCREASED INFORMATION NEED

Yearly indices give important information on the changes from one year to another. But yearly indices alone are insufficient, because the changes from one year to another are often caused by stochastic factors such as weather conditions, which are of little importance for nature conservation. It is more important to test whether systematic, overall changes (trends) occur over a number of years.

In addition, it is necessary to analyse the causes of trends, for instance the effects of human activities. Thus, in addition to providing yearly indices, a method is needed that can test for trends and is helpful in the analysis of possible causes.

INDEX METHOD

To cope with the demands mentioned, one of us (Jeroen Pannekoek) developed a new method and a user-friendly PC-computer program called TRIM (TRends and Indices for Monitoring data). The program computes indices and trends for time series by means of loglinear Poisson regression and takes the problems mentioned into account. Although such methods are also available in several standard computer packages (SPSS, GENSTAT, GLIM etc.), these are less well suited for the analysis of time series of several hundreds of sites together. To process such large data sets a faster algorithm has been developed than in the usual approach (Pannekoek & Van Strien, 1998).

Loglinear models are linear models for the logarithm of the expected counts. This is different from the more traditional approach to take the logarithm of the counts themselves, and can result in much better fitted models in case of many zero counts. The model can also be written as a multiplicative model for the untransformed expected counts. Here, we start with a model that describes yearly indices i.e. with effects for each year and each site. This model is closely related to those in the currently used methods of Mountford (1982) and Underhill & Prys-Jones (1994). The model can be expressed as

$$\mu_{ij} = a_i b_j \qquad (1)$$

where μ_{ij} is the expected count in site i and year j and $b_1 = 1$. The parameters a_i and b_j are the effects of site i and year j on the expected count. For year j the expected count ($a_i b_j$) is a factor b_j times the expected count ($a_i 1$) for the first year. The year parameters b_j can thus be regarded as yearly indices. Since the parameters b_j are independent of i, this model implies that the yearly indices are equal for each site.

A second model is a model with site effects and a multiplicative trend, written as

$$\mu_{ij} = a_i c^{(j-1)} \qquad (2)$$

which shows that for each site the expected count is a_i for the first year, a_ic for the second year, a_ic^2 for the third year etc. Model (2) can be viewed as a restricted version of model (1), with the restriction that $b_j = c^{(j-1)}$, showing that for this model the year parameters are equal for each year. It is possible to test this restriction, which means that we can test whether the year indices change with a constant factor or differ significantly from such behaviour.

Both model (1) and model (2) are rather restrictive because the yearly indices, b_j and c(j-1) respectively, are assumed to be the same for each site. By the use of covariants (variables that describe some property of the sites) this assumption can be relaxed and the models can be improved. For instance, if the sites can be classified according to habitat (e.g. woodland, farmland, dunes, heathland), we could apply model (1) to each habitat separately, thus arriving at a model with different yearly indices for sites classified as woodland, sites classified as farmland etc. Similarly, we could apply model (2) to each habitat separately, which results in a model with a different multiplicative trend for each habitat. It is possible to test for the significance of the effects of a covariant; that is to say we can test whether or not the yearly indices or trends differ significantly across the categories of a covariant. Apart from leading to improved estimates of indices and trends, covariants are also important for investigating possible causes of trends, such as human activities.

The method provides not only indices and trend estimates, but also standard errors of the parameter estimates, a Goodness-of-Fit test (deviance) for each model and a Wald test to test specific hypotheses. These tests are useful to select the models that fit best to the data. It is also possible to use weights to improve the estimates of national indices (see the example below for more details). Serial correlation is included, which affects the standard errors of the parameters and test results. Overdispersion can also be taken into account.

By using the computer program TRIM, one can select each of the models interactively and can chose whether or not to include serial correlation and/or overdispersion. The program produces a listing and optionally several result files: indices for each covariant category and fitted values per site. TRIM allows data up to 2 000 sites and 80 years and up to 10 covariants.

EXAMPLE

To illustrate the method, we used counts of 1984-1994 for the Woodlark *Lullula arborea* from the Breeding Bird Monitoring Program in The Netherlands. Nowadays, this scheme consists of more than 500 sample sites censused yearly, spread over the entire country. The sites are censused mostly by volunteers, using the territory mapping method (Van Dijk, 1996).

A number of models was estimated and tested, using 1984 as the base year and taking overdispersion and serial correlation into account. All models estimated took less than one minute on a Pentium computer. First, model (2) with one parameter for all years was estimated. This model fitted reasonably well, with a deviance of 452 ($df = 462$; $p = 0.63$). The numbers of Woodlark significantly increased with a factor 1.13 ($p = 0.02$), which implies an overall year-to-year increase of 13% (Fig. 1, upper line).

Instead of model indices, it is also possible to get "imputed indices" that are based on an imputed table. The table with imputed counts is formed from the table with observed counts, whereby missing values are replaced by model-based counts. The indices based on imputation are less model dependent than model-based indices because they depend on model-based estimates for the missing counts only. Therefore, imputed indices do not show a smooth curve like the model indices (Fig. 1).

Furthermore, model (1) with effects for each year was estimated (Fig. 1, lower line; deviance = 418; $df = 453$; $p = 0.88$). For this model, model-based indices are equal to imputed indices (see also Ter Braak et al., 1994). Testing the differences between model (1) and model (2) revealed that model (1) was a significant improvement over model (2) (Wald test = 37.0; $df = 9$; $p < 0.01$). In other words: indices based on model (1) are more plausible here than those based on model (2).

Figure 1. Indices for the number of breeding pairs of the Woodlark based on (1) a trend model, (2) a model whereby only missing values are based on a trend model and (3) a model with a parameter for each year (+). Number of sites involved = 116.

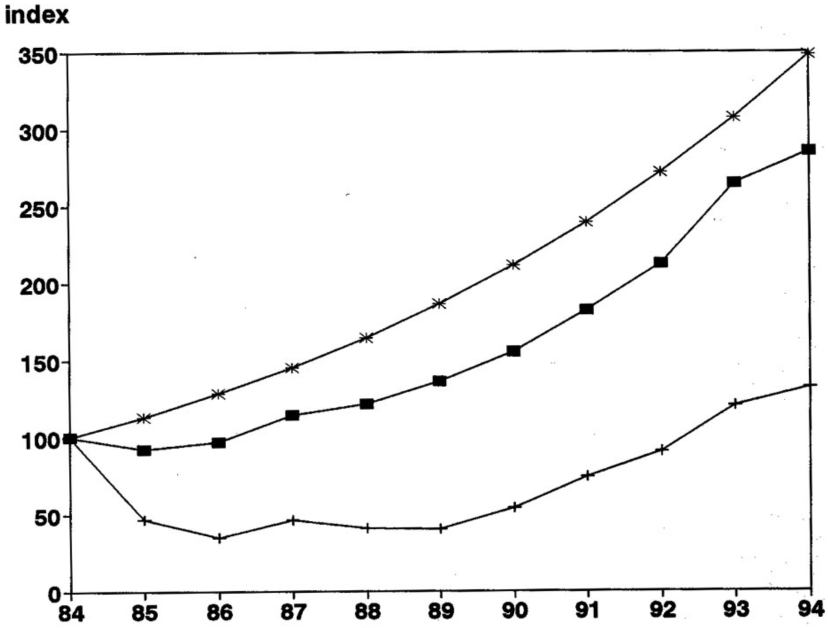
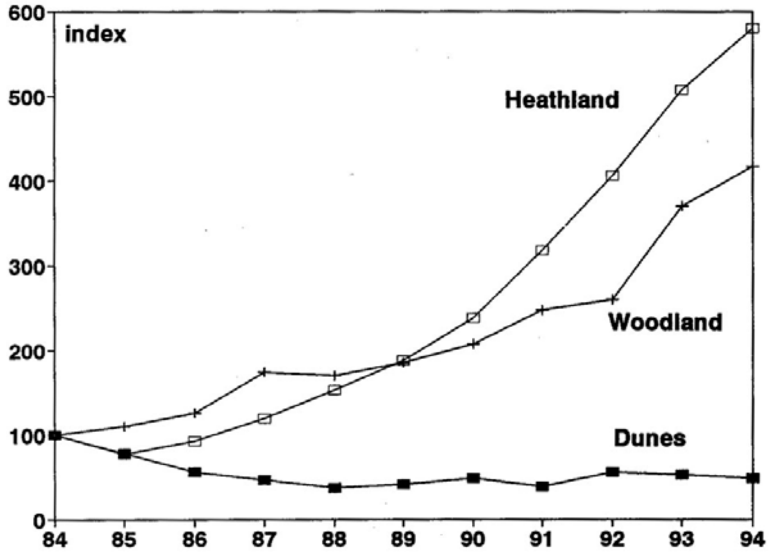


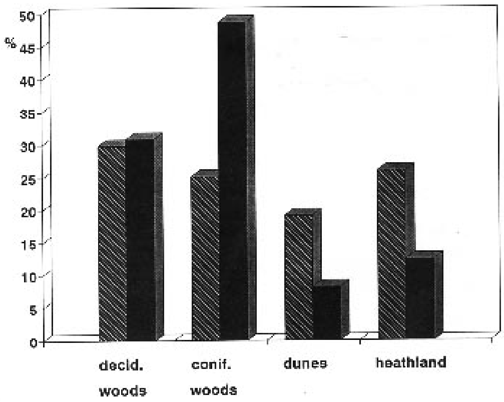
Figure 2: Indices for the number of breeding pairs of the Woodlark per habitat: (1) heathland, (2) woodland, whereby coniferous wood and deciduous woods were taken together, because indices were very similar and (3) dunes.



Thus, apart from assessing an estimate of the overall change across the entire period examined, we have assessed that the changes are definitely non-linear. The decrease from 1984 to 1985 and 1986 is due to severe winters, where after the species recovered and continuously increased from 1989 onwards (Van Dijk, 1990). This recent increase is remarkable because in many European countries this species seems to decrease or at least remains stable (Tucker et al., 1994).

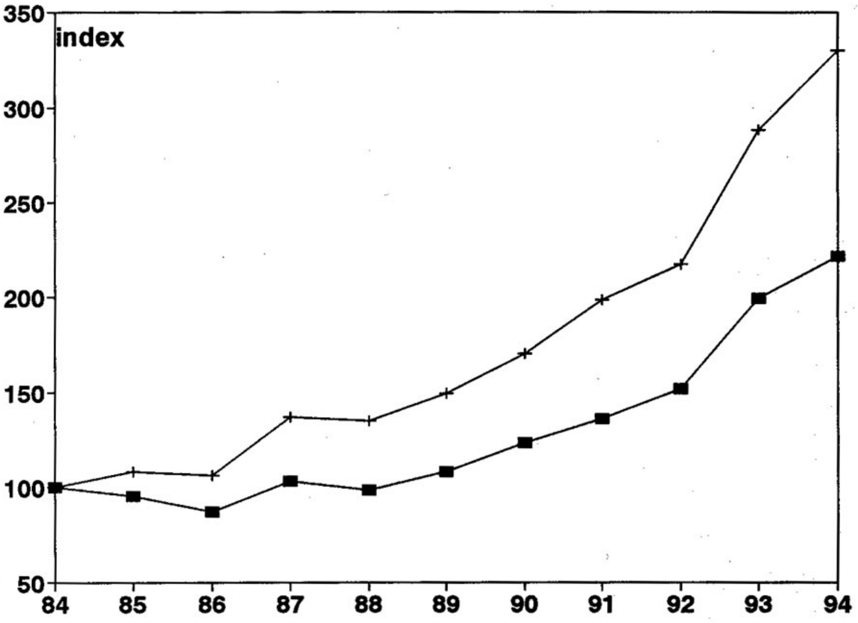
The Woodlark breeds in open habitats with few trees and bushes and occurs in dunes areas, heathland and forest clearings (Tucker et al., 1994). So far, we assumed that the same changes have occurred across all sites and thus do not differ between habitat types. To check this assumption, we applied a model with trends that may differ between habitat categories. The trend slope differed significantly between habitat categories (Wald test 38.7; *df* = 3; $p < 0.01$). The Woodlark has increased in heathland and woodland with a year-to-year change of 23 % and 16% respectively (Fig. 2). In coastal dune areas the numbers decreased with a yearly change of 6 %. This model fits better to the data than the model with one trend parameter for all sites. Due to lack of data, it was not possible to test a model with year effects that depend on habitat.

Figure 3: The distribution of the surface area of habitat categories in The Netherlands (left side) and in the Breeding Bird Monitoring Program (right side).



Because the trends differ between habitat types, we can not simply take all sites together to achieve a proper estimate of the national trend. The over- and undersampling of particular habitats may easily lead to incorrect national indices. By giving each observation a weight, we can adjust for the effects of over- and undersampling of particular habitats. Conifer plantations are undersampled and dunes and heathland are oversampled, as is shown by a comparison of the distribution of sites over habitat types in the Common Breeding Bird Census with the surface areas of these habitats in The Netherlands (Fig. 3). Adjustment for this phenomenon leads to other indices (Fig. 4). The lower line in Fig. 4 indicates the indices in case conifer plantations and dune areas are simply taken together. The upper line represents the trend if observations in sites in conifer plantations are given higher weights, and dune sites lower weights according to their surface areas. This results in a higher increase of this species, because of the lower contribution of the trends in dune sites, where the species has decreased most.

Figure 4: Indices for the number of breeding pairs of the Woodlark with (1) and without (2) adjustment for oversampling of dune areas and undersampling of coniferous woods. Heathland sites were left out of consideration here.



CONCLUSIONS AND FURTHER DEVELOPMENTS

Publications with results in terms of indices of species may easily suggest that there is one set of "true" yearly indices for each particular species. The example shows, however, that different models applied may lead to different results, some of them being more plausible than others. An important advantage of the loglinear method developed is that models can be tested against each other, so that better models can be selected, leading to, hopefully, more plausible indices.

So far, this method produces yearly indices and trends and is capable of assessing the effects of covariants on these changes, which is helpful in tracing the causes of changes. This method can also deal with difficulties inherent in the analysis of monitoring data, especially missing values, over- and undersampling of strata, serial correlation and violations of Poisson distribution assumption.

The method will be extended to analyse non-linear trends, which will become a necessity when the time series get longer. This leads to a model with several parameters describing the time series, instead of model (1) with a parameter for each year separately and model (2) with only one trend parameter for all years (see Pannekoek & Van Strien, 1998). We can also think of incorporating tests against a standard level, thus testing whether specified target levels have been achieved or not. This may be helpful in conservation practice and especially in the evaluation of policy measures. Another extension might be the adjustment of the time series for influences of factors like weather conditions, thereby enhancing the detection of trends (see also Van Strien et al., 1994).

REFERENCES

Crawford, T.J., 1991. The calculation of index numbers from wildlife monitoring data. In: F.B. Goldsmith (ed.). Monitoring for conservation and ecology, Chapman & Hall, London, p. 225-248.

Dobson, A.J., 1991. An introduction to generalized linear models. Wiley, New York.

Mountford, M.D., 1982. Estimation of population fluctuations with application to the Common Bird Census. Applied Statistics 31 (2):135-143.

McCullagh, P. & J.A. Nelder, 1989. Generalized Linear Models. 2nd edition. Chapman & Hall, London.

Pannekoek, J. & A. van Strien, 1998. TRIM 2.0 for Windows. (TRends & Indices for Monitoring data). Statistics Netherlands, Voorburg.

Ter Braak, C.J.F., A.J. van Strien, R. Meijer & T.J. Verstrael, 1994. Analysis of monitoring data with many missing values: which method? In: E.J.M. Hagemeijer & T.J. Verstrael (eds.), 1994. Bird Numbers 1992. Distribution, monitoring and ecological aspects. Proceedings of the 12th International Conference of IBCC and EOAC, Noordwijkerhout, The Netherlands. Statistics Netherlands, Voorburg/Heerlen & SOVON, Beek-Ubbergen, p. 663-673.

Tucker, G.M., M.F. Heath with L. Tomialojc & R.F.A. Grimmett, 1994. Birds in Europe: their conservation status. Birdlife Conservation Series no. 3. Birdlife International, Cambridge, p. 364-365.

Underhill, L.G. & R.P. Prys-Jones, 1994. Index numbers for waterbird populations. I: Review and methodology. J. Applied Ecology (31): 463-480.

Van Dijk, A.J. 1990. Strenge winters zetten de toon in de eerste vijf jaar BMP. Limosa 6: 141-152 (in Dutch with English summary).

Van Dijk, 1996. Broedvogels inventariseren in proefvlakken. Handleiding Monitoring Project (BMP). SOVON Vogelonderzoek Nederland, Beek-Ubbergen (in Dutch).

Van Strien, A.J., E.J.M. Hagemeijer & T.J. Verstrael, 1994. Estimating the probability of detecting trends in breeding birds: often overlooked but necessary. In: E.J.M. Hagemeijer & T.J. Verstrael (eds.), 1994. Bird Numbers 1992. Distribution, monitoring and ecological aspects. Proceedings of the 12th International Conference of IBCC and EOAC, Noordwijkerhout, The Netherlands. Statistics Netherlands, Voorburg/Heerlen & SOVON, Beek-Ubbergen, p. 525-531. Flade, M., 1992.